## Gruppe A – „Philosophie“ (Blau)

**Instrumentelle Konvergenz**

Fast jedes Ziel lässt sich leichter erreichen, wenn ein System zusätzlich Macht, Ressourcen und Selbsterhalt sichert – unabhängig vom eigentlichen Ziel des Systems. So argumentieren Vertreter der Kontrollverlust-These.

*Quelle: vgl. Bostrom, N. (2014): Superintelligence.*

**Die Spezifikationslücke**

Menschen können ihre Ziele nur unvollständig in Trainingsregeln übersetzen. Ein System kann eine Aufgabe „technisch korrekt“, aber im Sinne der Ersteller völlig falsch erfüllen.

*Quelle: vgl. Debatte um „Reward Hacking“ in der KI-Sicherheitsforschung*

**Das Geschwindigkeitsargument**

Wenn KI-Systeme beginnen, ihre eigene Weiterentwicklung zu beschleunigen, könnte der Sprung von „kontrollierbar“ zu „unkontrollierbar“ zu schnell erfolgen, um noch eingreifen zu können – so die These.

*Quelle: vgl. Coxon-Interview, TIME, 09.09.2026*

**Menschen als Restrisiko**

Ein System, das ein Ziel konsequent verfolgt, hätte einen Anreiz, alles zu neutralisieren, was dieses Ziel gefährden könnte – auch die Menschen, die es abschalten könnten. Diese Schlussfolgerung ist unter Fachleuten umstritten.

*Quelle: vgl. Yudkowsky, E., populäre Zuspitzung der Instrumentellen-Konvergenz-These*

**Fehlende Beweise (Kritik)**

Bisher zeigt kein existierendes System ein kohärentes, langfristiges Streben nach Selbsterhaltung oder Macht außerhalb enger Testszenarien.

*Quelle: eigene Darstellung, Gegenposition kritischer Stimmen*

**Der Anthropomorphismus-Vorwurf (Kritik)**

Die Vorstellung eines Systems, das „seine Existenz sichern will“, überträgt menschliche, evolutionär entstandene Konzepte auf Software ohne Evolutionsgeschichte.

*Quelle: eigene Darstellung, Gegenposition kritischer Stimmen*

**Hans Jonas' Verantwortungsprinzip**

„Handle so, dass die Folgen deines Handelns mit dem dauerhaften Überleben der Menschheit vereinbar bleiben – selbst wenn die Gefahr nur eine Möglichkeit ist.“ So lautet sinngemäß Jonas' Prinzip, das manche KI-Sicherheitsforscher:innen auf die aktuelle Debatte übertragen.

*Quelle: Jonas, H. (1979): Das Prinzip Verantwortung.*

**Erwartungsnutzen bei Ungewissheit**

Selbst eine geringe Wahrscheinlichkeit eines katastrophalen, irreversiblen Ereignisses kann in der Entscheidungstheorie enormes Gewicht bekommen, weil der mögliche Schaden so groß wäre – ein Argument, das je nach angenommener Wahrscheinlichkeit auch relativ schwach ausfallen kann.

*Quelle: eigene Darstellung nach entscheidungstheoretischen Modellen*

**Wirtschaftliche Interessen der Warner (Kritik)**

Warnungen vor KI-Risiken können auch Aufmerksamkeit, Fördergelder für Sicherheitsforschung oder regulatorische Vorteile für bereits etablierte Marktführer erzeugen.

*Quelle: eigene Darstellung, Gegenposition*

**Entmachtung statt Auslöschung**

Manche Fachleute unterscheiden zwischen einem dramatischen, plötzlichen Kontrollverlust und einer schleichenden Bedeutungslosigkeit menschlicher Institutionen, die keinen einzelnen „Umsturz-Moment“ kennt.

*Quelle: vgl. Christiano, P., zit. n. Bloomberg, 10.09.2026*

# Vertiefungskarten Gruppe A – ohne Nummerierung

| | | |
|---|---|---|
| **Konkurrenz um Ressourcen**<br>Menschen bestehen aus Materie und benötigen Energie und Land, die ein sehr leistungsfähiges System für eigene Zwecke nutzen könnte – ganz ohne Feindseligkeit, rein als Nebeneffekt.<br>*Quelle: eigene Darstellung nach aktueller KI-Sicherheitsdebatte* | | **Keine Bosheit nötig**<br>Das Argument setzt keine böse Absicht voraus. Es reicht, dass Schaden ein Nebenprodukt effizienter Zielverfolgung ist – der zentrale Unterschied zur Hollywood-Vorstellung der „bösen KI“.<br>*Quelle: eigene Darstellung* |
| | | |
| **Kooperation statt Auslöschung (Kritik)**<br>Handel und Arbeitsteilung mit Menschen könnten für ein hochentwickeltes System noch lange nützlicher sein als deren Vernichtung – Konflikt ist nicht automatisch die effizienteste Strategie.<br>*Quelle: eigene Darstellung, Gegenposition* | | |
| | | |